# Interregional and intraregional variability of intergroup attitudes predict online hostility

Rosenbusch, Hannes; Evans, Anthony; Zeelenberg, Marcel

*Published in:*
European Journal of Personality

*DOI:*
10.1002/per.2301

*Publication date:*
2020

*Document Version*
Publisher's PDF, also known as Version of record

*European Journal of Personality*, *Eur. J. Pers.* **34**: 859–872 (2020)
Published online 13 September 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2301

# Interregional and intraregional variability of intergroup attitudes predict online hostility

HANNES ROSENBUSCH*, ANTHONY M. EVANS and MARCEL ZEELENBERG
*Department of Social Psychology, Tilburg University, Tilburg, The Netherlands*

*Abstract: To what extent are intergroup attitudes associated with regional differences in online aggression and hostility? We test whether regional attitude biases towards minorities and their local variability (i.e. intraregional polarization) independently predict verbal hostility on social media. We measure online hostility using large US American samples from Twitter and measure regional attitudes using nationwide survey data from Project Implicit. Average regional biases against Black people, White people, and gay people are associated with regional differences in social media hostility, and this effect is confounded with regional racial and ideological opposition. In addition, intraregional variability in interracial attitudes is also positively associated with online hostility. In other words, there is greater online hostility in regions where residents disagree in their interracial attitudes. This effect is present both for the full resident sample and when restricting the sample to White attitude holders. We find that this relationship is also, in part, confounded with regional proportions of ideological and racial groups (attitudes are more heterogeneous in regions with greater ideological and racial diversity). We discuss potential mechanisms underlying these relationships, as well as the dangers of escalating conflict and hostility when individuals with diverging intergroup attitudes interact. © 2020 The Authors. European Journal of Personality published by John Wiley & Sons Ltd on behalf of European Association of Personality Psychology*

Key words: online hostility; regional dispositions; intergroup attitudes; attitude polarization; Twitter language

## INTRODUCTION

Hostile intergroup behaviour is frequently expressed through hateful speech on social media (Chau & Xu, 2007; Gerstenfeld, Grant, & Chiang, 2003). People use social media to express their outrage towards opposing groups (Crockett, 2017) and even endorse or threaten others with physical violence. Such instances of verbal hostility are facilitated through the anonymous nature of online environments, where aggression is less risky than it would be offline (see online disinhibition: Suler, 2004). Online aggression can occasionally spark offline violence, making online hostility a risk factor for both psychological and physical well-being (Hinduja & Patchin, 2007, 2012). This spill-over effect from online to offline aggression was documented in a series of studies by Müller and Schwarz (2019a, 2019b), who respectively used temporary social media outages and an instrumental variable strategy to ascertain the causal order of aggressive acts. Notice that the reverse effect, offline behaviour affecting online behaviour, has also been observed across a range of studies and that the order and interaction of both spheres are ongoing issues of debate (e.g. Greijdanus et al., 2020).

Importantly, online hostility often consists of inter*group* aggression, with minority members suffering from victimization more frequently than majority members (Awan & Zempi, 2016; Müller & Schwarz, 2019a). Online *hate* often does 'not attack individuals in isolation' but rather targets a collective of people (Hawdon, Oksanen, & Räsänen, 2017, p. 254). Psychological researchers therefore frequently measure vicarious experiences of online hate in which the reader is not personally attacked, but belongs to the derogated minority group (Tynes, Rose, & Williams, 2010). Given that online hostility towards minorities affects large amounts of people (Abbott, 2011; Costello, Hawdon, Ratliff, & Grantham, 2016), varies across geographic areas (Hawdon et al., 2017), and might even turn into physical violence (Awan & Zempi, 2016), it is important to identify the environments in which it is most likely to occur.

We analyse US American samples from Twitter and nationwide surveys from Project Implicit to examine geographical differences in online hostility. More precisely, we test whether regional *averages* of attitudinal biases towards minorities and their local *variability* (i.e. intraregional polarization) independently predict verbal hostility on social media. In the following section, we review prior research pointing to the idea that average regional attitudes towards minorities are related to regional differences in hostility. Subsequently, we argue that average regional attitudes do not tell the whole story and introduce the idea that it is also important to

*Correspondence to: Hannes Rosenbusch, Department of Social Psychology, Tilburg University, 5000 LE Tilburg, The Netherlands.
E-mail: h.rosenbusch@uvt.nl

This article earned Open Data and Open Materials badges through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The data and materials are permanently and openly accessible at https://osf.io/r69xj/. Author's disclosure form may also be found at the Supporting Information in the online version.



*Handling editor: John Rauthmann*
*Received 22 November 2019*
*Revised 27 July 2020, Accepted 2 August 2020*

consider how attitudinal biases are spread within local regions. More precisely, we propose that high variance in regional attitudes (which indicates the presence of conflicting ideological or demographic groups) is positively associated with online hostility.

## REGIONAL ATTITUDES TOWARDS MINORITIES AND ONLINE HOSTILITY

In recent years, online social media have become a major outlet for blatant intergroup discrimination (for reviews see Keum & Miller, 2018; Peterson & Densley, 2017). While offline discrimination often takes on subtle forms, online hostility is often blatant and explicit. Arguably, online aggression is bolstered by anonymity for perpetrators and decreased visibility of victims' suffering compared with offline settings (Kahn, Spencer, & Glaser, 2013). Accordingly, online abuse is common for both racial minorities (Tynes, Giang, Williams, & Thompson, 2008; Tynes, Reynolds, & Greenfield, 2004) and sexual minorities (Cooper & Blumenfeld, 2012; Varjas, Meyers, Kiperman, & Howard, 2013). In most cases, such forms of online hostility are argued to originate from antiminority biases including racism and homophobia, which vary across geographical locations (e.g. Hehman, Flake, & Calanchini, 2018; Swank, Frost, & Fahs, 2012).

What factors lead to the local emergence of hostile online environments? Online hostility can emerge from current local events (Williams & Burnap, 2015) or local history (Payne, Vuletich, & Brown-Iannuzzi, 2019). For example, Kaakinen, Oksanen, and Räsänen (2018) observed that the Paris terror attack from November 2015 was associated with a rise in fear and intergroup hostility among Finnish internet users, who related to the pre-attack situation of their fellow Europeans (c.f., Oksanen et al., 2018). According to the authors, this finding is in line with the general observation that threatening societal events serve as a trigger for outgroup blaming and intergroup hostilities. In their study, hostility was measured as the experienced frequency of verbal online hate. Groups that were targeted more often than before the event included religious, ethnic, and political groups, as these groups were labelled responsible for the attacks and the resulting societal uncertainty.

More generally, there is a long tradition of research rooted in the social identity approach on the connection between ingroup threat and outgroup hostility (Tajfel & Turner, 1979; Turner, 1985). Under threat, outgroup derogation appears to be a common strategy of regaining collective self-esteem (Branscombe & Wann, 1994). Specifically, regional experiences of ingroup threat seem to elicit heightened aggression and negative intergroup emotions (Fischer, Haslam, & Smith, 2010; Huddy & Feldman, 2011), which can be locally engrained and passed on over generations if the eliciting event was impactful enough (Obschonka et al., 2018; Payne et al., 2019). Note that regional construct aggregates (here, intergroup attitudes and hostility) are interpreted as a psychological facet of regional culture (Kitayama, Ishii, Imada, Takemura, & Ramaswamy, 2006) and that neither their interpretations nor their intercorrelations can be generalized to the individual level (Rentfrow, 2010; Rentfrow, Gosling, & Potter, 2008). Specifically, interracial biases aggregated on a regional level were defined as 'average, or collective, psychological predisposition' that local groups have towards each other (Hehman, Calanchini, Flake, & Leitner, 2019, p. 1025).

In the context of group relations, local animosity is associated with negative phenomena for all the involved groups. For instance, the more negatively White locals feel towards Black (compared with White) people, the more stress-related health problems Black locals experience (specifically circulatory diseases) and the higher the mortality for both Black and White locals rises (Leitner, Hehman, Ayduk, & Mendoza-Denton, 2016a, 2016b; Orchard & Price, 2017). Similarly, regional levels of anti-Black sentiment predict lethal police force against Black people (Hehman et al., 2018). Arguably, racially biased environments put a strain on both intergroup and interpersonal relationships, thereby enforcing the local propensity for violence against minority members (Hehman et al., 2018). Similarly, Johnson and Chopik (2019) found that in US counties with strong racial stereotypes, the targeted minority group itself also engages in more violence (measured as rates of murder, aggravated assault, and illegal weapon possession). Thus, there appears to be a relationship between average regional attitudes and local rates of violent behaviours. This relationship is in line with common correlations between aggregated psychological measurements and local indices of well-being (Plaut, Markus, & Lachman, 2002).

Importantly, the link between average antiminority attitudes and regional hostility holds true for other types of (e.g. nonracial) intergroup attitudes. There is substantial regional variation in attitudes towards other minority groups including gay men and lesbian women (gay people from here). While regional covariates of attitudes towards gay people have received less attention, prior research suggests that the well-being of sexual minorities is linked to their regional environments (Morandini, Blaszczynski, Dar-Nimrod, & Ross, 2015). For example, gay people living in southern states of the USA were subjected to increased levels of discrimination and stigma (e.g. thinly veiled hostility) compared with gay people living in non-southern states (Swank et al., 2012).

Building on previous research, we investigate if biases against minorities and prevalence of verbal online hostility are associated across geographical spaces. However, we go beyond prior work, which focused on comparing *average* regional attitudes, to investigate if hostility is related to the *distribution* of intergroup attitudes within regions. We test if high levels of hostility are more likely to be observed in regions where individuals with conflicting attitudes and ideologies are more likely to come into contact. In other words, hostility may also be observed in regions with high levels of intraregional *variability* in attitudes.

## INTRAREGIONAL ATTITUDE VARIABILITY AND ONLINE HOSTILITY

While average local attitudes (i.e. the extent to which citizens from a region are, on average, biased against certain groups) are an important indicator of regional culture, they paint a simplified picture and discard valuable environmental information. Hostility is usually the result of people having divergent, rather than convergent, attitudes (Harinck & Ellemers, 2014), and local averages do not capture the level of local divergence. Intergroup conflicts emerge when two groups *disagree* about adequate group hierarchy (often involving their own group; Bobo, 1999). That is to say, compared with the effects of *average* levels of social biases, intraregional *variability* of social biases may be more strongly associated with regional hostility. Variability in relative attitudes towards minorities implies ideological opposition or disagreement between the different attitude holders (e.g. some are pro-White, whereas others are pro-Black). Past work highlights that such ideological opposition can indeed lead to substantial aggression and segregation (Brandt, Crawford, & Van Tongeren, 2019; Brandt, Reyna, Chambers, Crawford, & Wetherell, 2014; Kouzakova, Ellemers, Harinck, & Scheepers, 2012).

Donald Trump's presidential candidacy in 2016 illustrates how diverging intergroup attitudes can create a hostile online environment. Arguably, dehumanizing and aggressive antiminority rhetoric led to hostile backlashes among targeted minority members (Kteily & Bruneau, 2017) and among majority members whose egalitarian attitudes were in dissonance with Trump's antiminority standpoint (Meyer & Tarrow, 2018). These recent examples suggest that ideological polarization, especially regarding minority treatment, often entails aggression (Iyengar & Westwood, 2015; Miller & Conover, 2015). This aggression is frequently expressed through partisan activity on social media (Crockett, 2017; Hasell & Weeks, 2016), which in turn often crosses the threshold of hate speech (e.g. Ben-David & Matamoros-Fernández, 2016). Theoretical concepts like perceived injustice (van Zomeren, Postmes, & Spears, 2008), status instability (Scheepers, 2009), and politicized identities (van Zomeren, Postmes, & Spears, 2012) all point towards conflict under diverging attitudes (rather than consensually positive or negative attitudes).

In sum, regions not only differ in the average attitudes towards minorities (see previous section), but also in how variable these attitudes are within each region (Evans & Need, 2002). In regions with high attitude variability, biased people clash with people who hold more egalitarian values (or with people who are biased in favour of minority groups). Given that polarized attitudes towards minorities spur conflict and hostility, divided regions should be characterized by frequent aggression, especially in anonymous online environments. Conversely, if locals are homogenous and similarly biased (i.e. there is little attitude variability), then there is less reason for local conflict (see Figure 1 for two example counties).

Our central hypothesis is that intraregional *variability* in minority attitudes is correlated with online hostility. Importantly, we expect this correlation to remain significant even after controlling for the effect of average regional attitudes. In other words, we expect that variance in social biases (i.e. regional heterogeneity) is positively associated with online hostility.

Importantly, intergroup attitudes are likely to covary according to regional group composition. Two opposing groups of people, each preferring their ingroup, imply relatively polarized intergroup attitudes and therefore reason for intergroup hostility (Tajfel & Turner, 1979). Given the importance of social identity for eliciting intergroup biases, we assume that the effect of attitude variability on local hostility will be subdued when controlling for regional differences in racial and political diversity. That is, opposition between opinion holders is likely confounded with opposition between racial or political groups. Social identity research suggests that local contact between groups with opposing group attitudes (e.g. racial or political groups) can spark anxiety and intergroup tension (Tausch, Hewstone, Kenworthy, Cairns, & Christ, 2007; Zeitzoff, 2017). Notice that this proposed connection between local opposition and hostility appears to contradict the very influential contact hypothesis, which states that intergroup contact should, under certain conditions, improve intergroup relations (Allport, 1954). This apparent theoretical contradiction has been treated by multiple scholars observing that regional diversity in the USA constitutes a special case (Rae, Newheiser, & Olson, 2015; van der Meer & Tolsma, 2014). Most importantly, local outgroup presence/diversity in this context often does not entail personal, beneficial intergroup interaction, which could suppress negative threat effects (for a detailed discussion see Laurence, 2014). Therefore, introducing local diversity into predictive models should subtract from the effect of attitude variability on hostility.

## METHOD

We examine how social media hostility is associated with regional averages and regional variability in relative attitudes

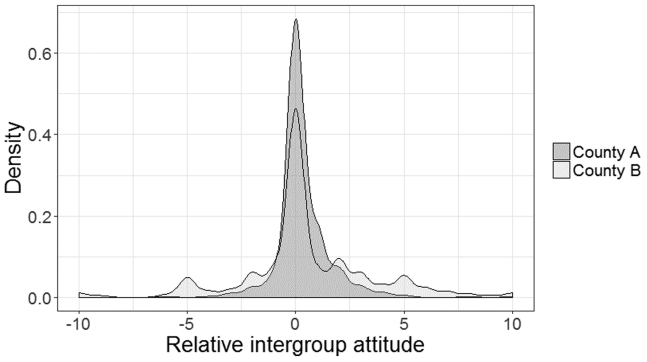


Figure 1. Counties A and B have somewhat similar mean levels of social bias ($M_A$ = 0.33 and $M_B$ = 0.51; percentiles 18 and 38), whereas they clearly differ in terms of intraregional variability. The distribution of bias in county A is less variable (*SD* = 1.26, percentile 1) than the distribution of bias in county B (*SD* = 3, percentile 99). In other words, members of county A are more homogenous in their intergroup attitudes than members of county B.

towards minorities. Specifically, we focus on US American counties and examine (i) attitudes towards Black people relative to attitudes towards White people (attitudinal bias) and (ii) attitudes towards gay people (men and women) relative to attitudes towards heterosexual people. We focus on these groups because attitudes towards minorities (including Black people and gay people) are a prevalent topic in US American politics and social media discussions. We use two operationalizations of attitude variability (standard deviation and kurtosis). We use large social media samples (from Twitter) to measure regional differences in hostility, as online language reflects regional variations in psychological phenomena (Eichstaedt et al., 2015). We include two measures of social media hostility (expressions of anger and swearing). All data and code for the current work can be found online (https://osf.io/r69xj/). We refer to the supporting information by pointing out the specific files in question.

## Sample

In many US counties, both the Project Implicit and Twitter datasets have zero or very few measurements, which does not allow for meaningful county-level scores. Thus, researchers have to decide how many measurements are sufficient to compute a county-level score and include the county in subsequent analyses. In the past, researchers have applied different cut-off scores. We selected US counties with at least 172 racial attitude scores per county, which leads to a sample of 1094 counties, as this constitutes the average of prior research using a range of different cut-offs (Leitner et al., 2016a, 2016b; ; Orchard & Price, 2017; Rae et al., 2015). We excluded one additional county from Alaska (FIPS 02110), because it complicated our corrections for spatial autocorrelation (it is ~900 km away from the next county). The remaining 1093 counties were selected to have sufficient individual attitude scores to assure the reliability of our aggregated county-level attitude measures. About 2,000 counties or similar small regions are not included in the sample because of insufficient local measurements. Figure 2 depicts the geographical coverage of the utilized sample.

In order to examine the effect of our cut-off decision, we conducted sensitivity analyses with samples of 815 and 1280 counties (at least 301 and 124 scores per county, respectively); these alternative cut-offs corresponded to the minimum and maximum sample sizes used previously for the

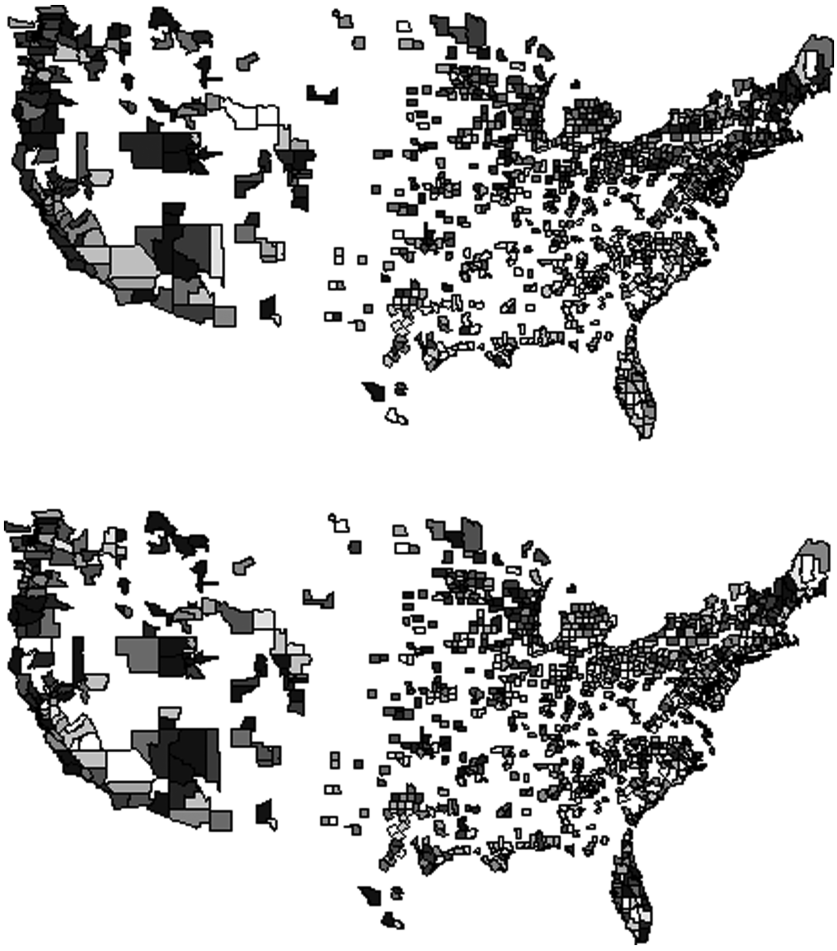

Figure 2. Top: Attitude variability scores (in deciles) of US counties in the utilized sample. Lighter tones indicate a higher variability. Bottom: Relative frequency of anger expression (in deciles) of US counties in the utilized sample. Lighter tones indicate higher frequencies of anger expression on Twitter. For additional maps showing the distribution of average bias, bias towards gay people, and frequency of swearing, please see the folder 'maps' in the Open Science Framework (OSF) repository.

racial attitudes sample (Leitner et al., 2016a; Leitner et al., 2016b; Orchard & Price, 2017; Rae et al., 2015). In our sensitivity analyses, only 4% of all tests for the full and the White population failed to replicate [average absolute beta coefficient deviation of 0.029; see file 'ethnicity bias (main analyses for all & white residents).R' in the supporting information], indicating the robustness of our primary results. However, results for Black respondents were highly sensitive to sample restrictions, with 39% of tests varying in their conclusions (average beta coefficient deviation of 0.035; see file 'ethnicity bias (all analyses for black residents).R'). This instability was likely the outcome of the limited availability of data for the county-level attitude measure (i.e. there were often few individual attitude scores from Black respondents). Thus, results for this group should be interpreted with caution. Unless specified otherwise, we always report the most conservative result in the main text.

There is less prior research using data on regional attitudes towards gay people, so we did not use prior research cut-offs. Instead, we used counties that had at least as many attitude scores as the county with the smallest number of scores for the racial attitude data above (172 scores). Thus, for the analyses on relative attitudes towards sexual minorities, we restricted our analyses to 677 counties. In line with the analyses for racial attitudes, we again conducted sensitivity analyses with limits of at least 301 and 124 scores per county, respectively. The sensitivity of the results was again quite high for the relative attitudes towards gay people as 25% of test results differed in the sensitivity analyses (average beta coefficient deviation of 0.034). We again advise to interpret the results for relative attitudes towards sexual minorities with caution. All conclusions drawn from the analyses were compatible with the results in the main text and the sensitivity analyses.

## Measures

All measures were on the county level. We used explicit attitude measures, as the meaning of implicit test scores on individual and collective levels remains uncertain (Blanton & Jaccard, 2017). However, the utilized data source for explicit attitude scores also contain implicit measures (Project Implicit; Xu, Nosek, & Greenwald, 2014), and we include implicit measures in the supporting information (see folder 'unprocessed data from past publications'). Selection biases are a potential limitation of relying on data from Project Implicit, meaning some resident groups (e.g. women, young people, Xu et al., 2014, and educated people, Morris & Ashburn-Nardo, 2009) are overrepresented while others are underrepresented. This problem is present in virtually all large datasets (e.g. Gosling Potter Internet Project, Gosling, Vazire, Srivastava, & John, 2004; BBC Lab dataset, Rentfrow et al., 2013). In order to obtain more representative county scores, we therefore employed raking by age, gender, and education. Raking (for an introduction, see Battaglia, Izrael, Hoaglin, & Frankel, 2009) is the process of comparing one's sample to a representative sample (often census data of the target population) on a range of, usually demographic, variables. If the demographic distribution in one's sample deviates from the target population, say there are more women in the sample, the scores of the underrepresented group, in this case men, receive larger weights when computing aggregate scores (say the sample mean or standard deviation of a variable). Thus, in the current project, we compared the demographics of each county's Project Implicit sample with the county's census data (US Census Bureau, 2017; US Department of Agriculture, 2017) and reweighted participant scores, so that county-level attitude scores (mean, standard deviation, and kurtosis) are more closely in line with the expected population scores. Hoover and Dehghani (2019) provide a discussion of sample representativeness in large subnational datasets.[1] All variables were standardized for better comparability of the results. Notice that attitude scores were collected between 2003 and 2017 while online hostility was measured between 2009 and 2010. This temporal overlap prevents claims of one-directional causality (as does the correlational nature of the data).

### *Interracial attitudes*

The data were obtained from OSF repositories (https://osf.io/52qxl/) and are described by Xu et al. (2014). We estimated the regional level (mean) and variability (standard deviation) of bias in interracial attitudes using geo-tagged scores of warmth felt towards Black people on an 11-point scale subtracted from warmth felt towards White people. Thus, positive scores indicate a relative pro-White bias, a score of 0 indicates a neutral attitude, and negative scores indicate a pro-Black bias. Given that the two individual warmth questions were presented back-to-back in the survey and that they were formatted in the same way (the only difference being the target group), we assume that participants were very conscious of the difference between their two answers and that this numerical difference can be interpreted as explicit bias. Previous research therefore generally utilized this operationalization of explicit bias (e.g. Connor, Sarafidis, Zyphur, Keltner, & Chen, 2019; Hehman et al., 2018; Leitner et al., 2016a; Leitner et al., 2016b; Payne et al., 2019). The difference scores computed for each participant were subsequently used to compute two scores per county: the average local difference score (interpreted as regional bias) and the standard deviation of the local difference scores (interpreted as local disagreements in bias). Validation studies of the utilized data and measurements are described by Hehman et al. (2019). The sample included attitude measurements from 2 048 781 participants (county-level minimum = 172, median = 624, maximum = 54 235). Separate analyses are added for the White and Black subsamples (respectively 1 634 117 and 283 239 participants).

### *Attitudes towards gay and straight people*

The Project Implicit data can be obtained from Project Implicit's OSF repositories (https://osf.io/ajdgr). We estimated relative attitudes towards gay people using geo-tagged scores

[1]The supporting information includes an earlier version of our analyses conducted without raking; note that this procedure does not substantively change our findings.

of warmth felt towards gay men and lesbian women on an 11-point scale subtracted from warmth felt towards heterosexual men and women (Xu et al., 2014). The relative attitude scores for men and women were averaged into one score indicating attitudes towards gay people relative to attitudes towards heterosexual people. Thus, a higher score indicates a stronger bias against gay (or in favour of heterosexual) people. The sample included attitude measurements from 763 907 participants (county-level minimum = 172, median = 513, maximum = 22 412).

*Verbal online hostility*

Regional variation in online hostility was assessed through Twitter language extracted from US counties. The dataset (provided by Eichstaedt et al., 2015) included 148 000 000 tweets and was successfully used in the original publication to make psychological comparisons between US counties. We utilized the LIWC2015 software (Pennebaker, Boyd, Jordan, & Blackburn, 2015) to count swear words (e.g. 'bullshit') and amount of anger expressions (e.g. 'annoyed', 'angry', and 'stupid'), with relative frequencies being interpreted as regional levels of hostility on social media. The LIWC measures of anger and swearing were used previously for assessing hostility (Hancock, Woodworth, & Boochever, 2018; Ksiazek, 2015; Matsumoto, Hwang, & Frank, 2016). While we believe that the validation procedures for the hostility measures should generally be conducted on the individual level (see citations), we ascertain their validity on the collective level in the supporting information (see file 'validity of hostility measure.R').

*Racial and political proportions*

We obtained the regional numbers of Black and White residents from the website of the US Census Bureau (2012–2016 data, 2017) and the amount of votes for Donald Trump versus Hillary Clinton from McGovern (2017).

### Analysis plan

In the results section, we analyse the association between racial attitudes and regional hostility. Our primary analyses consisted of linear regression analyses in which we computed the effects of regional attitudes and attitude variability on social media hostility. First, we probed average attitudes among all residents, then we conducted separate analyses using average attitudes among only White and only Black residents as predictors of hostility. Note, however, that we always focused on overall regional hostility as our dependent variable, as the county-level Twitter dataset did not allow us to differentiate between the hostility of White versus Black Twitter users.

For each analysis, we first report the results for a simple regression of regional hostility on average attitude bias. Then, we introduce regional proportions of racial and ideological groups as covariates into the same model. After our analyses on average levels of regional bias, we introduce variability in attitudes as an additional predictor of regional hostility. Here, we again present results for all, White, and Black residents, and under inclusion of the additional covariates. We also dedicate one section to attitude kurtosis as an alternative measure to the dispersion of attitudes. Lastly, we replicate the analyses for relative attitudes towards gay people. In the supporting information, we include additional analyses with post-hoc county matching on further covariates (county-level income, employment, crime rates) to ascertain the observed effects described in the main text. All effects of attitude variability were robust in these analyses (see file 'matched controls.R').

### Assumption checks

The assumptions of heteroscedastic and normally distributed errors were checked through residual plots. Residual maps and a significant Moran's $I$ statistic indicated that the assumption of independent observations was violated through spatial autocorrelation (see file 'test spatial autocorrelation. R'). That means counties in close vicinity of each other had similar model residuals. To account for this, we added an autocovariate to each regression model, which used the weighted average of hostile online language in neighbouring counties as predictors. This correction substantially improved our satisfaction with the residual maps. However, in some cases, the Moran's $I$ was still statistically significant at $\alpha = .05$. As it is always greatly reduced, as maps no longer show visible patterns, as and the inclusion of the autocovariate does not seem to shift our effects of interest, we assume that the remaining autocorrelation does not threaten the conclusions drawn from the data. For brevity, we do not describe results for the highly significant autocovariate term for each model, but they are included in the supplementary results (see all. R files in the folder 'results presented in the main text').[2]

## RESULTS

We report descriptive statistics for the primary variables in Table 1. Counties were, on average, biased against Black people, $t(1{,}092) = 57.322$, $p < .001$, and gay people, $t(676) = 61.867$, $p < .001$.

Table 1. Descriptive statistics

| Variable | *M* | *SD* | Min | Max |
|---|---|---|---|---|
| Average attitude (White-Black) | 0.575 | 0.331 | −1.016 | 1.704 |
| Variability attitude (White-Black) | 2.100 | 0.340 | 1.122 | 3.463 |
| Average attitude (straight-gay) | 1.193 | 0.502 | −0.266 | 3.380 |
| Variability attitude (straight-gay) | 2.748 | 0.265 | 1.901 | 4.031 |
| Anger | 0.003 | 0.0006 | 0.0008 | 0.007 |
| Swearing | 0.002 | 0.0008 | 0.0007 | 0.008 |

*Note*: The mean scores of anger/swearing can be interpreted as follows: '0.3% of all tweeted words in the county were expressions of anger'. All listed (nonstandardized) variables were standardized before the analyses.

### Average racial attitudes, regional variance, and online hostility

In the following section, we test how the average interracial attitudes of local residents relate to local online hostility. In the second paragraph of the section, we change the focus from average attitudes to the variability in attitudes.

### Analyses including attitudes of all residents

Average anti-Black (pro-White) attitudes (averaged across *all* local citizens) were associated with decreased swearing ($\beta = -0.145$, 95% confidence interval (CI) [−0.199, −0.091], $p < .001$). This effect was rendered nonsignificant when controlling for local proportions of Black and White people ($\beta = -0.048$, 95% CI [−0.113, 0.017], $p = .149$). For anger, the effect was also negative and nonsignificant ($\beta = -0.027$, 95% CI [−0.082, 0.029], $p = .344$). Thus, there is no strong evidence that counties with different levels of average anti-Black bias show different levels of online hostility.

In contrast to county-level means, county-level variability in bias was positively associated with anger ($\beta = 0.268$, 95% CI [0.212, 0.323], $p < .001$) and swearing ($\beta = 0.327$, 95% CI [0.269, 0.384], $p < .001$), even after controlling for average regional attitudes, race proportions, and ideological proportions (the proportions of Trump and Clinton votes in the 2016 election, $\beta_{anger} = 0.171$, 95% CI [0.106, 0.236], $p < .001$; $\beta_{swearing} = 0.231$, 95% CI [0.168, 0.294], $p < .001$). Introducing racial and political control variables decreased the size of the bias variability slopes by an average of 32.8% (see Figure 3). In the full model, the relative number of Clinton voters (over Trump voters) was associated with less online hostility ($\beta_{anger} = -0.201$, 95% CI [−0.271, −0.131], $p < .001$; $\beta_{swearing} = -0.171$, 95% CI [−0.237, −0.105], $p < .001$) while racial diversity was associated with more online hostility ($\beta_{anger} = 0.710$, 95% CI [0.460, 0.960], $p < .001$; $\beta_{swearing} = 0.854$, 95% CI [0.616, 1.092], $p < .001$).

### Separate analyses for White and Black attitude holders

In this section, we repeat the first set of analyses, but instead of computing the mean and standard deviation of attitudes of *all* residents, we compute them separately for White and Black residents. This allows us to test whether average attitudes and variability in attitudes show the same relationship with hostility across racial groups. When restricting attitude measurements to White residents, average anti-Black bias positively predicted hostility ($\beta_{anger} = 0.228$, 95% CI [0.174, 0.283], $p < .001$; $\beta_{swearing} = 0.201$, 95% CI [0.146, 0.257], $p < .001$). Conversely, when restricting attitude measurements to Black residents, *pro*-Black (anti-White) bias marginally predicted hostility, with the absolute effect size being much smaller than for White residents and not significant for the anger measure ($\beta_{anger} = -0.055$, 95% CI [−0.111, 0.001], $p = .054$; $\beta_{swearing} = -0.063$, 95% CI [−0.117, −0.009], $p = .023$). In other words, hostility levels were high in counties where White residents had strong anti-Black biases and where Black residents had strong anti-White biases. Simultaneously introducing local proportions of racial and ideological groups into the models renders these effects nonsignificant for White residents ($\beta_{anger} = 0.060$, 95% CI [−0.024, 0.143], $p = .162$; $\beta_{swearing} = 0.020$, 95% CI [−0.061, 0.102], $p = .620$) and Black residents ($\beta_{anger} = -0.038$, 95% CI [−0.097, 0.021],

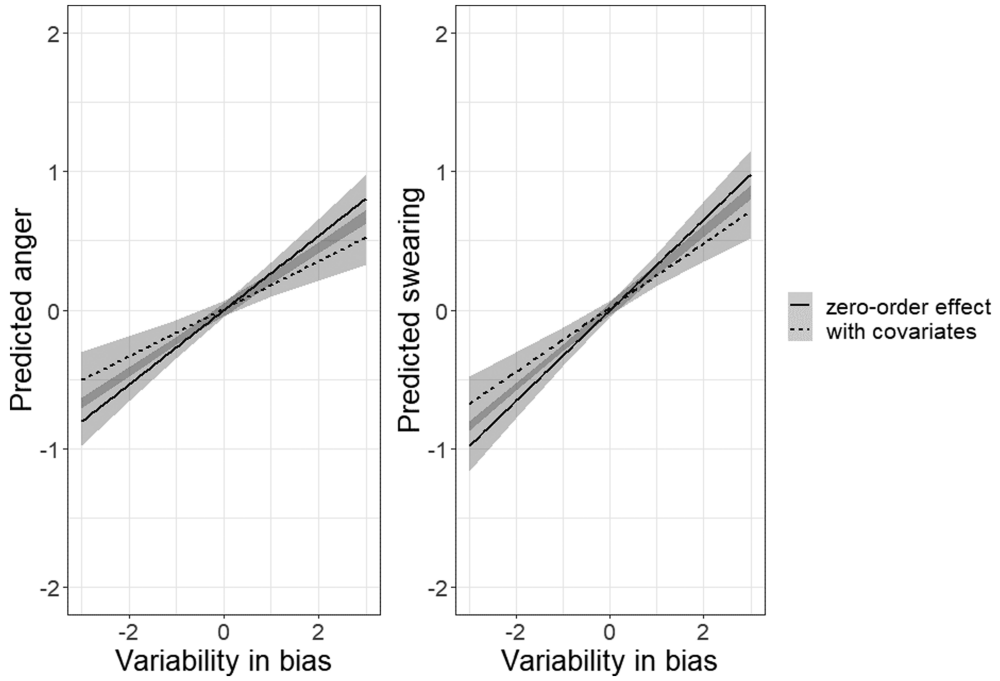


Figure 3. The association between bias variability and regional hostility. Local levels of hostility were positively associated with intraregional variability in bias. Covariates are average bias, and local racial and ideological proportions.

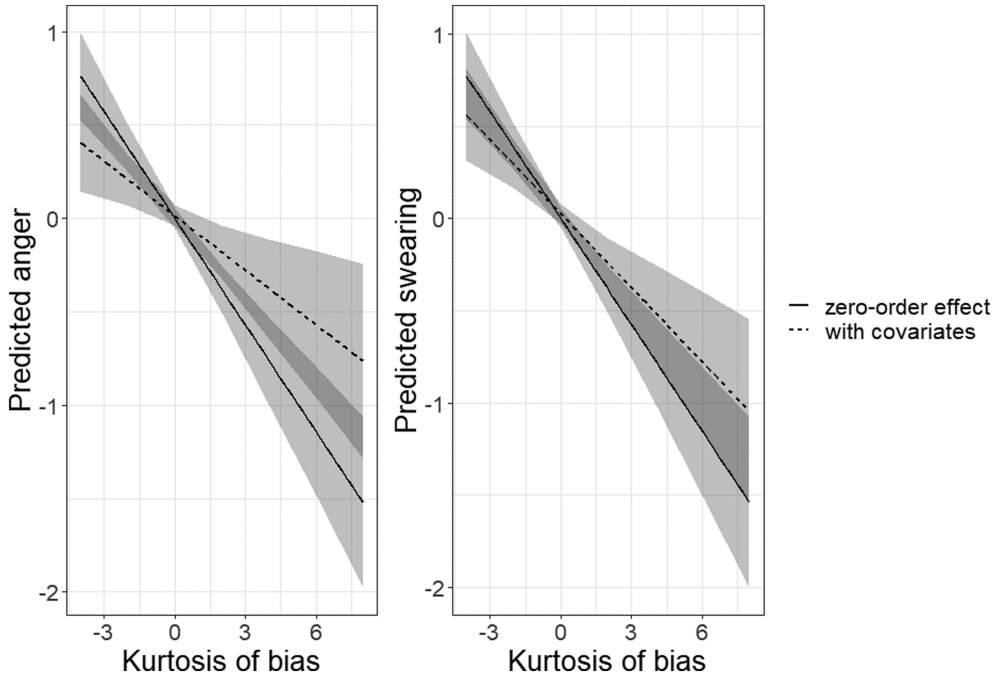


Figure 4. The association between bias kurtosis and regional hostility. The kurtosis of attitudes, serving as a reversed measure of polarization, was negatively associated with hostility. Thus, polarization again predicted hostility. Covariates are average bias, and racial and ideological proportions.

$p$ = .209; $\beta_{swearing}$ = −0.045, 95% CI [−0.100, 0.010], $p$ = .106).

County-level variability in attitudes among White residents emerged as a positive predictor of online hostility ($\beta_{anger}$ = 0.221, 95% CI [0.167, 0.275], $p$ < .001; $\beta_{swearing}$ = 0.210, 95% CI [0.162, 0.271], $p$ < .001). Thus, diversity in interracial attitudes among White residents is also associated with online hostility. Again, these effects are substantially decreased when introducing local proportions of racial and ideological groups into the model. For white residents, the decrease in effect size was 45.3% ($\beta_{anger}$ = 0.104, 95% CI [0.026, 0.183], $p$ = .009; $\beta_{swearing}$ = 0.131, 95% CI [0.055, 0.206], $p$ < .001). County-level variability in attitudes among Black residents was not significantly associated with online hostility ($\beta_{anger}$ = −0.008, 95% CI [−0.064, 0.047], $p$ = .774; $\beta_{swearing}$ = −0.007, 95% CI [−0.061, 0.047], $p$ = .805).

## Analyses using kurtosis as a measure of regional polarization

In our previous analyses, we estimated attitudinal variability within each region using the standard deviation of regional intergroup attitudes. As a higher standard deviation implies a larger dispersion of attitudes, this operationalization is in line with our reasoning. However, the relationship between regional standard deviations and attitudinal *polarization* is merely indirect. High polarization means that many cases score at the extremes of the distribution. Clearly, this phenomenon can contribute to high standard deviations, but a more direct measure of polarization is a variable's kurtosis. The (Pearson) kurtosis does *not*, as often assumed, quantify the peakedness of a distribution, but tail extremity (i.e. the propensity of extreme values on either side of the distribution; Westfall, 2014). Smaller kurtosis values indicate saturated tails with relatively many values close to the poles, which makes kurtosis a good (negative) measure of polarization. Using a sample's kurtosis to assess polarization of intergroup attitudes is supported by the work of DiMaggio, Evans, and Bryson (1996), who argued that kurtosis is likely better suited than standard deviation to assess polarization. The following two subsections replicate all previous effects of attitude variability, but operationalize the construct as attitude kurtosis rather than standard deviation of attitudes.

### *Analyses including attitudes of all residents*

When replacing attitude variability in the upper analyses with bias kurtosis, all full-sample effects of kurtosis regardless of hostility measure or covariates are statistically significant (all $\beta$s ≤ −0.097, all $p$s ≤ .003; see Figure 4).[3]

Regions with higher levels of kurtosis (indicating fewer extreme scores) had less online hostility compared with regions with lower levels of kurtosis (indicating more extreme scores and greater regional polarization). The inclusion of local proportions of racial and political groups as covariates decreased the initial effect size by 46.9%. As attitudinal polarization can be expected to relate closely to both standard deviation and kurtosis of opinions (the measures are correlated $r$ = −.535), it is not surprising that the results from earlier replicate to such a large degree (see Table 2 for a side-by-side view).

[3]When loosening sample restriction to 1280 counties and controlling for average bias, regional racial opposition, and regional ideological opposition, the attitude kurtosis is not significantly associated with swearing ($\beta$ = −0.057, $p$ = .058).

Table 2. Predicting anger and swearing with regional attitude variability and kurtosis

| Models predicting anger | *Attitude variability* | *Attitude kurtosis* |
|---|---|---|
| Simple regression | $\beta = 0.268$, $p < .001$ | $\beta = -0.190$, $p < .001$ |
| Including average attitudes + covariates | $\beta = 0.171$, $p < .001$ | $\beta = -0.097$, $p = .003$ |
| **Models predicting swearing** | | |
| Simple regression | $\beta = 0.327$, $p < .001$ | $\beta = -0.193$, $p < .001$ |
| Including average attitudes + covariates | $\beta = 0.231$, $p < .001$ | $\beta = -0.106$, $p = .004$ |

*Note*: Attitude variability refers to the standard deviation of attitudes.

*Separate analyses for White and Black attitude holders*

When restricting the sample of attitude holders to White locals, the attitude kurtosis again predicted hostility ($\beta_{anger} = -0.167$, 95% CI [−0.222, −0.112], $p < .001$; $\beta_{swearing} = -0.155$, 95% CI [−0.210, −0.099], $p < .001$). When simultaneously controlling for the above-mentioned covariates, the kurtosis of attitudes no longer predicted verbal hostility ($\beta_{anger} = -0.033$, 95% CI [−0.105, 0.039], $p = .368$; $\beta_{swearing} = -0.050$, 95% CI [−0.120, 0.019], $p = .156$; drop in effect size 74%). For Black residents, we also found significant effects of attitude kurtosis on online hostility ($\beta_{anger} = -0.082$, 95% CI [−0.138, −0.027], $p = .004$; $\beta_{swearing} = -0.068$, 95% CI [−0.122, −0.014], $p = .014$).[4] Again, when simultaneously controlling for the above-mentioned covariates, the kurtosis of attitudes no longer predicted verbal hostility ($\beta_{anger} = -0.050$, 95% CI [−0.106, 0.006], $p = .080$; $\beta_{swearing} = -0.030$, 95% CI [−0.082, 0.022], $p = .255$; drop in effect size 47.5%). This pattern of results for the full, White, and Black sample is consistent with the analyses above, with the exception that local proportions of racial and ideological groups actually accounts for all the variance explained by attitude kurtosis among White residents.

### Relative attitudes towards gay people and online hostility

In this last section, we used the above approach to examine the association between relative attitudes towards gay people and online hostility. This allowed us to investigate whether intergroup attitudes towards other minorities, and the local distribution of these attitudes, also predicted hostility on social media.

The average anti-gay attitude (i.e. averaged across *all* local citizens) positively predicted hostility ($\beta_{anger} = 0.320$, 95% CI [0.252, 0.388], $p < .001$; $\beta_{swearing} = 0.321$, 95% CI [0.251, 0.391], $p < .001$). When introducing attitude variability, ideological proportions, and racial proportions as covariates, the effect of average attitudes on anger was no longer significant ($\beta_{anger} = 0.100$, 95% CI [−0.012, 0.205], $p = .081$; $\beta_{swearing} = 0.121$, 95% CI [0.016, 0.227], $p = .025$).[5] Variability in relative attitudes towards gay people also positively predicted regional levels of hostility ($\beta_{anger} = 0.307$, 95% CI [0.237, 0.376], $p < .001$; $\beta_{swearing} = 0.287$, 95% CI [0.215, 0.358], $p < .001$). When entering the covariate set into the model, the effect of attitude variability on swearing became nonsignificant ($\beta_{anger} = 0.134$, 95% CI [0.037, 0.231], $p = .007$; $\beta_{swearing} = 0.061$, 95% CI [−0.033, 0.156], $p = .204$; drop in effect size 67.6%). When replacing the standard deviation of attitudes with its kurtosis, the regional attitude kurtosis positively predicted online hostility ($\beta_{anger} = -0.237$, 95% CI [−0.306, −0.167], $p < .001$; $\beta_{swearing} = -0.225$, 95% CI [−0.297, −0.153], $p < .001$). When entering regional divides as covariates into the model, the effect of attitude kurtosis decreased by 60.1% ($\beta_{anger} = -0.098$, 95% CI [−0.178, −0.019], $p = .016$; $\beta_{swearing} = -0.086$, 95% CI [−0.164, −0.009], $p = .029$[6]). We refrained from estimating the effect of relative attitudes towards gay people held by people with specific sexual orientations, as county-level effects become increasingly unstable with lower numbers of attitude scores. However, the data in the supporting information allow for such analyses. For a full list of all stepwise inferential tests, please see code/results in the folder 'results presented in the main text'.

## DISCUSSION

We set out to test the relationship between regional biased attitudes towards minority groups and regional levels of verbal hostility on Twitter. The present analysis shows multiple connections between both phenomena, supporting past research on the broad spectrum of negative correlates of regional attitudinal bias. First, we found that average levels of anti-Black bias were not reliably associated with online hostility (i.e. the significance and magnitude of the relationship depended on the operationalization of hostility and the covariate set). However, these results differed (and became clearer) when we examined the separate attitudes of White versus Black county residents. Average levels of anti-Black bias among White residents were positively associated online hostility, whereas the opposite relationship was observed when viewing the attitudes of Black residents. In other words, we observed greater hostility in counties where White residents held stronger pro-White biases and in counties where Black residents held stronger pro-Black biases. Thus, anti-*outgroup* attitudes were positively associated with hostility, although not beyond the underlying effects of local racial and ideological group proportions.

[4]These effects are not significant when loosening sample restrictions to 1280 counties ($\beta_{anger} = 0.023$, $p = .375$; $\beta_{swearing} = -0.032$, $p = .208$)

[5]The effect on swearing was also no longer significant when applying either looser ($\beta = 0.088$, $p = .071$) or tighter ($\beta = 0.103$, $p = .104$) sample restrictions.

[6]In fact, the effects became nonsignificant when applying either looser ($\beta_{anger} = -0.071$, $p = .065$; $\beta_{swearing} = -0.046$, $p = .209$) or tighter ($\beta_{anger} = -0.075$, $p = 0.118$; $\beta_{swearing} = -0.078$, $p = .095$) sample restrictions.

Critically, attitudinal variability was more strongly associated with online hostility than regional attitude averages. Overall, regions with dispersed racial attitudes were more hostile compared with regions with less attitudinal variability. This effect was present in the total sample, as well as in the analyses restricted to White residents (whereas the effect was less stable among Black residents). Further, the introduction of regional divides between racial and voter groups subtracts from the effect of attitude variability. In other words, controlling for the presence of conflicting ideology or demographic groups reduces the relationship between attitudinal variability and hostility, sometimes to a point where attitude variability is no longer a significant predictor. This pattern of results suggests that the observed effects might be primarily because of tensions between racial and ideological groups, whose ingroup identities are closely attached to interracial relations (Leach & Allen, 2017; Perry & Whitehead, 2015).

When conducting similar analyses using relative attitudes towards gay people, the main effect of average anti-gay bias on regional hostility was significant, but did not remain significant when controlling for regional divides (as the relative presence of Trump supporters appears to be closely aligned with regional attitudes towards gay people). Results also differed regarding the effect of attitude variability. It was distinctly weaker than in the analyses of interracial attitudes, and could almost always be fully accounted for by controlling for ideological divides and average attitudes towards gay people. While it is true that attitude variability measures were less reliable and effects were less stable across sensitivity analyses compared with the analyses of interracial attitudes, we assume that findings might just not be generalizable across minority groups. While discrimination of either kind remains a divisive issue, racism might be a more frequent cause of hostility given the larger size and visibility of racial minority groups, and the USA's regionally specific history with slavery. Thus, racial attitudes might relate to regional online conflict relatively more often than homophobia or simply contribute more to regional stress and tacit intergroup tension.

### Intergroup conflict and online hostility

Computational and social identity research has described relationships between ideological opposition, group conflict, and aggressive behaviour on social media (Bail et al., 2018; Cicchirillo, Hmielowski, & Hutchens, 2015; Kwon & Cho, 2017; Kwon & Gruzd, 2017; Ott, 2017; Postmes, Spears, Sakhel, & De Groot, 2001; Spears & Postmes, 2015). We find evidence extending this research line by showing that regions with relatively variable interracial biases, as largely captured in local proportions of racial and political groups, are characterized by more hostility on social media compared to other regions.

Scholars have argued that the USA is experiencing an increase in ideological polarization (Twenge, Honeycutt, Prislin, & Sherman, 2016) and that treatment of minorities poses one of the most divisive topics today (Schaffner, MacWilliams, & Nteta, 2018). This phenomenon occasionally makes for an explosive mix when combined with the often-discussed online disinhibition effect, which describes people expressing more anger and hatred online than they would in person (e.g. Suler, 2004). Crockett (2017, p. 717) stated that 'Polarization in the US is accelerating at an alarming pace, with widespread and growing declines in trust and social capital. If digital media accelerates this process further still, we ignore it at our peril'. While the current research does not address whether this dynamic is in fact cascading over time, we revealed that ideological divisions and social media hostility are associated across geographical regions.

We want to highlight that while we find correlational evidence for a connection between attitude variability and online hostility, the nonexperimental data and the temporal overlap of variable measurements prevent identification of causal mechanisms. Multiple phenomena are likely responsible for the correlation between regional attitude variability and social media hostility. An obvious candidate explanation is that people disagree with other users holding different attitudes, for instance, by expressing outrage and insulting each other, which should be more likely to occur in areas where anti-Black sentiment clashes with egalitarian or anti-White sentiment. Another explanation is that the social uncertainty resulting from divided neighbourhoods is expressed through negative affect and venting online. Both directions from prejudice to online hostility (Bliuc, Faulkner, Jakubowicz, & McGarty, 2018) and from online hostility to prejudice (e.g. through desensitization; Soral, Bilewicz, & Winiewski, 2018) have been suggested in previous psychological research and both are in line with the observed effects as well as the decrease in effect sizes when introducing regional divides. Lastly, it is possible that the presence of certain extremist groups contributes to both the variability in bias and habitual anger/swearing online. Aggressive online activity of such extremist groups is, in turn, likely to spark similarly emotional backlash by opponents.

While the preceding mechanisms have a common root, opposition sparking hostility, there is another intriguing explanation for the statistical association. Perpetrators (and victims; Kaakinen, Keipi, Oksanen, & Räsänen, 2018) of online hate tend to have certain dispositions (Kurek, Jose, & Stuart, 2019; McCreery & Krach, 2018), which can be clustered across geographical areas (for psychological traits of US regions, see Rentfrow et al., 2013). It is reasonable to assume that a region's dispositional hostility can foster polarization and cross-group rejection. Imagine intergroup relations in an area where people are prone to swear and curse at each other. This is to say that the causal directionality of ideological variability and hostility might well be reversed. We estimate that a bidirectional relationship is most likely with polarization and online hostility mutually enforcing each other as an explosive mix. Crockett (2017, p. 771) might have forecasted this finding by asking 'If moral outrage is a fire, is the internet like gasoline?'

### Limitations

In the current work, we could only interpret the hostility measure as an all-inclusive county-level characteristic. While social biases were suggested to heighten hostility for both perpetrator and victim (Borders & Hennebry, 2015; Weber, Lavine, Huddy, & Federico, 2014), it would be interesting to examine who was hostile towards whom. While it is possible to employ predictive methods to estimate demographic information from individual Twitter profiles (Kteily, Rocklage, McClanahan, & Ho, 2019), the linguistic dataset we worked with only includes text aggregated over many anonymous users thereby preventing such methods. Other shortcomings are related to spatial nature of the utilized datasets. For instance, the well-known modifiable areal unit problem (Manley, 2014) applies to the current work. This problem suggests that aggregated measures can be biased by both the shape and the scale of the unit of analysis (counties). Our focus on *county*-level differences was based on convention, rather than theoretical justification. Future research should consider whether the present results would replicate on a city or even neighbourhood level. More fine-grained analyses would allow a more detailed examination of local intergroup relations and online hostility. Relatedly, despite the large amounts of data available through Project Implicit and the Twitter publication, the geographical coverage is far from complete, as indicated in Figure 2. This has been an issue throughout all past spatial analyses of the datasets and future efforts to fill blind spots through targeted data collection or value imputation would be highly beneficial.

## CONCLUSION

We find that a wide intraregional variation in relative attitudes towards minorities is associated with hostility on social media. This pattern is seemingly stronger when examining attitudinal bias towards racial, rather than sexual, minorities. The effect of intraregional variability runs parallel to regional divides between racial groups and political groups; and controlling for the local proportions of these groups reduces the association between intraregional variability and online hostility. Together, the results suggest that ideological polarization is accompanied with local unrest and aggression on social media. Further research is needed to pinpoint the dynamic processes that give rise to this association.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Data S1.** Supporting Information

## REFERENCES


Abbott, M. K. (2011). *Cyberbullying experiences of ethnic minorities*. La Verne, California: University of La Verne.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., … Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*, 9216–9221. https://doi.org/10.1073/pnas.1804840115

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Blanton, H., & Jaccard, J. (2017). You can't assess the forest if you can't assess the trees: Psychometric challenges to measuring implicit bias in crowds. *Psychological Inquiry*, *28*, 249–257. https://doi.org/10.1080/1047840X.2017.1373550

Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75–86. https://doi.org/10.1016/j.chb.2018.05.026

Bobo, L. D. (1999). Prejudice as group position: Microfoundations of a sociological approach to racism and race relations. *Journal of Social Issues*, *55*, 445–472. https://doi.org/10.1111/0022-4537.00127

Borders, A., & Hennebry, K. A. (2015). Angry rumination moderates the association between perceived ethnic discrimination and risky behaviors. *Personality and Individual Differences*, *79*, 81–86. https://doi.org/10.1016/j.paid.2015.01.047

Brandt, M. J., Crawford, J. T., & van Tongeren, D. R. (2019). Worldview conflict in daily life. *Social Psychological and Personality Science*, *10*, 35–43. https://doi.org/10.1177/1948550617733517

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, *23*, 27–34. https://doi.org/10.1177/0963721413510932

Branscombe, N. R., & Wann, D. L. (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology*, *24*, 641–657. https://doi.org/10.1002/ejsp.2420240603

Battaglia, M. P., Izrael, D., Hoaglin, D. C., & Frankel, M. R. (2009). Practical considerations in raking survey data. *Survey Practice*, *2*(5), 1–10.

Chau, M., & Xu, J. (2007). Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies*, *65*, 57–70. https://doi.org/10.1016/j.ijhcs.2006.08.009

Cicchirillo, V., Hmielowski, J., & Hutchens, M. (2015). The mainstreaming of verbally aggressive online political behaviors. *Cyberpsychology, Behavior, And Social Networking*, *18*, 253–259. https://doi.org/10.1089/cyber.2014.0355

Connor, P., Sarafidis, V., Zyphur, M. J., Keltner, D., & Chen, S. (2019). Income inequality and White-on-Black racial bias in the United States: Evidence from project implicit and Google trends. *Psychological Science*, *30*, 205–222. https://doi.org/10.1177/0956797618815441

Cooper, R. M., & Blumenfeld, W. J. (2012). Responses to cyberbullying: A descriptive analysis of the frequency of and impact on LGBT and allied youth. *Journal of LGBT Youth*, *9*, 153–177. https://doi.org/10.1080/19361653.2011.649616

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*, 311–320. https://doi.org/10.1016/j.chb.2016.05.033

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*, 769–771. https://doi.org/10.1038/s41562-017-0213-3

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized? *American Journal of Sociology*, *102*, 690–755. https://doi.org/10.1086/230995

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*, 159–169. https://doi.org/10.1177/0956797614557867

Evans, G., & Need, A. (2002). Explaining ethnic polarization over attitudes towards minority rights in Eastern Europe: A multilevel analysis. *Social Science Research*, *31*, 653–680. https://doi.org/10.1016/S0049-089X(02)00018-2

Fischer, P., Haslam, S. A., & Smith, L. (2010). "If you wrong us, shall we not revenge?" Social identity salience moderates support for retaliation in response to collective threat. *Group Dynamics: Theory, Research, and Practice*, *14*, 143–150. https://doi.org/10.1037/a0017970

Gerstenfeld, P. B., Grant, D. R., & Chiang, C. P. (2003). Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy*, *3*, 29–44. https://doi.org/10.1111/j.1530-2415.2003.00013.x

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies?: A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Greijdanus, H., de Matos Fernandes, C. A., Turner-Zwinkels, F., Honari, A., Roos, C. A., Rosenbusch, H., & Postmes, T. (2020). The psychology of online activism and social movements: Relations between online and offline collective action. *Current Opinion in Psychology*, *35*, 49–54. https://doi.org/10.1016/j.copsyc.2020.03.003

Hancock, J. T., Woodworth, M., & Boochever, R. (2018). Psychopaths online: The linguistic traces of psychopathy in email, text messaging and Facebook. *Media and Communication*, *6*, 83–92. https://doi.org/10.17645/mac.v6i3.1499

Harinck, F., & Ellemers, N. (2014). How values change a conflict. In C. K. W. De Dreu (Ed.), *Social conflict within and between groups* (pp. 19–36). Hove, UK: Psychology Press.

Hasell, A., & Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Human Communication Research*, *42*, 641–661. https://doi.org/10.1111/hcre.12092

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*, 254–266. https://doi.org/10.1080/01639625.2016.1196985

Hehman, E., Calanchini, J., Flake, J. K., & Leitner, J. B. (2019). Establishing construct validity evidence for regional measures of explicit and implicit racial bias. *Journal of Experimental Psychology: General*, *148*, 1022–1044. https://doi.org/10.1037/xge0000623

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*, 393–401. https://doi.org/10.1177/1948550617711229

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, *6*, 89–112. https://doi.org/10.1300/J202v06n03_06

Hinduja, S., & Patchin, J. W. (2012). Cyberbullying: Neither an epidemic nor a rarity. *European Journal of Developmental Psychology*, *9*, 539–543. https://doi.org/10.1080/17405629.2012.706448

Hoover, J., & Dehghani, M. (2019). The big, the bad, and the ugly: Geographic estimation with flawed psychological data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000240

Huddy, L., & Feldman, S. (2011). Americans respond politically to 9/11: Understanding the impact of the terrorist attacks and their aftermath. *American Psychologist*, *66*, 455–467. https://doi.org/10.1037/a0024894

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*, 690–707. https://doi.org/10.1111/ajps.12152

Johnson, D. J., & Chopik, W. J. (2019). Geographic variation in the Black-violence stereotype. *Social Psychological and Personality Science*, *10*, 287–294. https://doi.org/10.1177/1948550617753522

Kaakinen, M., Keipi, T., Oksanen, A., & Räsänen, P. (2018). How does social capital associate with being a victim of online hate? Survey evidence from the United States, the United Kingdom, Germany, and Finland. *Policy & Internet*, *10*, 302–323. https://doi.org/10.1002/poi3.173

Kaakinen, M., Oksanen, A., & Räsänen, P. (2018). Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, *78*, 90–97. https://doi.org/10.1016/j.chb.2017.09.022

Kahn, K. B., Spencer, K., & Glaser, J. (2013). Online prejudice and discrimination: From dating to hating. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding our online behavior* (2nd ed., pp. 201–219). New York: Oxford University Press 10.1093/acprof:oso/9780199639540.003.0011.

Kitayama, S., Ishii, K., Imada, T., Takemura, K., & Ramaswamy, J. (2006). Voluntary settlement and the spirit of independence: Evidence from Japan's "northern frontier". *Journal of Personality and Social Psychology*, *91*, 369–384. https://doi.org/10.1037/0022-3514.91.3.369

Ksiazek, T. B. (2015). Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. *Journal of Broadcasting & Electronic Media*, *59*, 556–573. https://doi.org/10.1080/08838151.2015.1093487

Keum, B. T., & Miller, M. J. (2018). Racism on the internet: Conceptualization and recommendations for research. *Psychology of Violence*, *8*, 782–791. https://doi.org/10.1037/vio0000201

Kouzakova, M., Ellemers, N., Harinck, F., & Scheepers, D. (2012). The implications of value conflict: How disagreement on values affects self-involvement and perceived common ground. *Personality and Social Psychology Bulletin*, *38*, 798–807. https://doi.org/10.1177/0146167211436320

Kteily, N., & Bruneau, E. (2017). Backlash: The politics and real-world consequences of minority group dehumanization. *Personality and Social Psychology Bulletin*, *43*, 87–104.

Kteily, N. S., Rocklage, M. D., McClanahan, K., & Ho, A. K. (2019). Political ideology shapes the amplification of the accomplishments of disadvantaged vs. advantaged group members. *Proceedings of the National Academy of Sciences*, *116*, 1559–1568. https://doi.org/10.1073/pnas.1818545116

Kurek, A., Jose, P. E., & Stuart, J. (2019). 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Computers in Human Behavior*, *98*, 31–40. https://doi.org/10.1016/j.chb.2019.03.027

Kwon, K. H., & Cho, D. (2017). Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review*, *35*, 84–102. https://doi.org/10.1177/0894439315602664

Kwon, K. H., & Gruzd, A. (2017, January). *Is aggression contagious online? A case of swearing on Donald Trump's campaign videos on YouTube*. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Laurence, J. (2014). Reconciling the contact and threat hypotheses: Does ethnic diversity strengthen or weaken community

inter-ethnic relations? *Ethnic and Racial Studies*, *37*, 1328–1349. https://doi.org/10.1080/01419870.2013.788727

Leach, C. W., & Allen, A. M. (2017). The social psychology of the Black Lives Matter meme and movement. *Current Directions in Psychological Science*, *26*, 543–547. https://doi.org/10.1177/0963721417719319

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016a). Blacks' death rate due to circulatory diseases is positively related to Whites' explicit racial bias: A nationwide investigation using Project Implicit. *Psychological Science*, *27*, 1299–1311. https://doi.org/10.1177/0956797616658450

Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016b). Racial bias is associated with ingroup death rate for Blacks and Whites: Insights from Project Implicit. *Social Science and Medicine*, *170*, 220–227. https://doi.org/10.1016/j.socscimed.2016.10.007

Manley, D. (2014). Scale, aggregation, and the modifiable areal unit problem. In M. M. Fischer, & P. Nijkamp (Eds.), *Handbook of regional science* (pp. 1157–1171). Berlin, Germany: Springer 10.1007/978-3-642-23430-9_69.

Matsumoto, D., Hwang, H. C., & Frank, M. G. (2016). The effects of incidental anger, contempt, and disgust on hostile language and implicit behaviors. *Journal of Applied Social Psychology*, *46*, 437–452. https://doi.org/10.1111/jasp.12374

McCreery, M. P., & Krach, S. K. (2018). How the human is the catalyst: Personality, aggressive fantasy, and proactive-reactive aggression among users of social media. *Personality and Individual Differences*, *133*, 91–95. https://doi.org/10.1016/j.paid.2017.06.037

McGovern, T. (2017). *US president county-level election results for 2012 and 2016*. Retrieved from https://github.com/tonmcg/County_Level_Election_Results_12-16

Meyer, D. S., & Tarrow, S. (Eds) (2018). *The resistance: The Dawn of the anti-Trump opposition movement*. New York: Oxford University Press.

Miller, P. R., & Conover, P. J. (2015). Red and blue states of mind: Partisan hostility and voting in the United States. *Political Research Quarterly*, *68*, 225–239. https://doi.org/10.1177/1065912915577208

Morandini, J. S., Blaszczynski, A., Dar-Nimrod, I., & Ross, M. W. (2015). Minority stress and community connectedness among gay, lesbian and bisexual Australians: A comparison of rural and metropolitan localities. *Australian and New Zealand Journal of Public Health*, *39*, 260–266. https://doi.org/10.1111/1753-6405.12364

Morris, K. A., & Ashburn-Nardo, L. (2009). The implicit association test as a class assignment: Student affective and attitudinal reactions. *Teaching of Psychology*, *37*, 63–68.

Müller, K., & Schwarz, C. (2019a). *Fanning the flames of hate: Social media and hate crime*. SSRN. https://doi.org/10.2139/ssrn.3082972

Müller, K., & Schwarz, C. (2019b). *From hashtag to hate crime: Twitter and anti-minority sentiment*. *SSRN*. https://doi.org/10.2139/ssrn.3149103

Obschonka, M., Stuetzer, M., Rentfrow, P. J., Shaw-Taylor, L., Satchell, M., Silbereisen, R. K., … Gosling, S. D. (2018). In the shadow of coal: How large-scale industries contributed to present-day regional differences in personality and well-being. *Journal of Personality and Social Psychology*, *115*, 903–927. https://doi.org/10.1037/pspp0000175

Oksanen, A., Kaakinen, M., Minkkinen, J., Räsänen, P., Enjolras, B., & Steen-Johnsen, K. (2018). Perceived societal fear and cyberhate after the November 2015 Paris terrorist attacks. *Terrorism and Political Violence*, *32*, 1–20.

Orchard, J., & Price, J. (2017). County-level racial prejudice and the Black–White gap in infant health outcomes. *Social Science and Medicine*, *181*, 191–198. https://doi.org/10.1016/j.socscimed.2017.03.036

Ott, B. L. (2017). The age of Twitter: Donald J. trump and the politics of debasement. *Critical Studies in Media Communication*, *34*, 59–68. https://doi.org/10.1080/15295036.2016.1266686

Payne, B. K., Vuletich, H. A., & Brown-Iannuzzi, J. L. (2019). Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, *116*, 11693–11698.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works*, *2015*, 1–22. https://doi.org/10.15781/T29G6Z

Perry, S. L., & Whitehead, A. L. (2015). Christian nationalism and white racial boundaries: Examining whites' opposition to interracial marriage. *Ethnic and Racial Studies*, *38*, 1671–1689. https://doi.org/10.1080/01419870.2015.1015584

Peterson, J., & Densley, J. (2017). Cyber violence: What do we know and where do we go from here? *Aggression and Violent Behavior*, *34*, 193–200. https://doi.org/10.1016/j.avb.2017.01.012

Plaut, V. C., Markus, H. R., & Lachman, M. E. (2002). Place matters: Consensual features and regional variation in American well-being and self. *Journal of Personality and Social Psychology*, *83*, 160–184. https://doi.org/10.1037/0022-3514.83.1.160

Postmes, T., Spears, R., Sakhel, K., & De Groot, D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, *27*, 1243–1254. https://doi.org/10.1177/01461672012710001

Rae, J. R., Newheiser, A. K., & Olson, K. R. (2015). Exposure to racial out-groups and implicit race bias in the United States. *Social Psychological and Personality Science*, *6*, 535–543. https://doi.org/10.1177/1948550614567357

Rentfrow, P. J. (2010). Statewide differences in personality: Toward a psychological geography of the United States. *American Psychologist*, *65*, 548–558. https://doi.org/10.1037/a0018194

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, *105*, 996–1012. https://doi.org/10.1037/a0034434

Rentfrow, P. J., Gosling, S. D., & Potter, J. (2008). A theory of the emergence, persistence, and expression of geographic variation in psychological characteristics. *Perspectives on Psychological Science*, *3*, 339–369. https://doi.org/10.1111/j.1745-6924.2008.00084.x

Schaffner, B. F., Macwilliams, M. C., & Nteta, T. (2018). Understanding white polarization in the 2016 vote for president: The sobering role of racism and sexism. *Political Science Quarterly*, *133*, 9–34. https://doi.org/10.1002/polq.12737

Scheepers, D. (2009). Turning social identity threat into challenge: Status stability and cardiovascular reactivity during inter-group competition. *Journal of Experimental Social Psychology*, *45*, 228–233. https://doi.org/10.1016/j.jesp.2008.09.011

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*, 136–146. https://doi.org/10.1002/ab.21737

Spears, R., & Postmes, T. (2015). Group identity, social influence, and collective action online. In S. Shyam Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 23–46). Oxford, UK: John Wiley & Sons.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*, 321–326. https://doi.org/10.1089/1094931041291295

Swank, E., Frost, D. M., & Fahs, B. (2012). Rural location and exposure to minority stress among sexual minorities in the United States. *Psychology & Sexuality*, *3*, 226–243. https://doi.org/10.1080/19419899.2012.700026

Tajfel, H., & Turner, J. C. (1979). An integrative theory of inter-group conflict. In W. G. Austin, & S. Worchel (Eds.), *The*

*social psychology of intergroup relations* (pp. 33–47). Monterey, CA: Brooks-Cole.

Tausch, N., Hewstone, M., Kenworthy, J., Cairns, E., & Christ, O. (2007). Cross-community contact, perceived status differences, and intergroup attitudes in Northern Ireland: The mediating roles of individual-level versus group-level threats and the moderating role of social identification. *Political Psychology*, *28*, 53–68. https://doi.org/10.1111/j.1467-9221.2007.00551.x

Turner, J. C. (1985). Social categorization and the self-concept: A social cognitive theory of group behavior. In E. J. Lawler (Ed.), *Advances in group processes: Theory and research* (pp. 77–122). Greenwich, CT: JAI.

Twenge, J. M., Honeycutt, N., Prislin, R., & Sherman, R. A. (2016). More polarized but more independent: Political party identification and ideological self-categorization among US adults, college students, and late adolescents, 1970–2015. *Personality and Social Psychology Bulletin*, *42*, 1364–1383. https://doi.org/10.1177/0146167216660058

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online racial discrimination and psychological adjustment among adolescents. *Journal of Adolescent Health*, *43*, 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

Tynes, B., Reynolds, L., & Greenfield, P. M. (2004). Adolescence, race, and ethnicity on the Internet: A comparison of discourse in monitored vs. unmonitored chat rooms. *Journal of Applied Developmental Psychology*, *25*, 667–684. https://doi.org/10.1016/j.appdev.2004.09.003

Tynes, B. M., Rose, C. A., & Williams, D. R. (2010). The development and validation of the online victimization scale for adolescents. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *4*, 2.

US Department of Agriculture (2017). *County-level data sets*. Retrieved from https://www.ers.usda.gov/data-products/county-level-data-sets

US Census Bureau (2017). *Population estimates*. Retrieved from https://factfinder.census.gov

van der Meer, T., & Tolsma, J. (2014). Ethnic diversity and its effects on social cohesion. *Annual Review of Sociology*, *40*, 459–478. https://doi.org/10.1146/annurev-soc-071913-043309

van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, *134*, 504–535. https://doi.org/10.1037/0033-2909.134.4.504

van Zomeren, M., Postmes, T., & Spears, R. (2012). On conviction's collective consequences: Integrating moral conviction with the social identity model of collective action. *British Journal of Social Psychology*, *51*, 52–71. https://doi.org/10.1111/j.2044-8309.2010.02000.x

Varjas, K., Meyers, J., Kiperman, S., & Howard, A. (2013). Technology hurts? Lesbian, gay, and bisexual youth perspectives of technology and cyberbullying. *Journal of School Violence*, *12*, 27–44. https://doi.org/10.1080/15388220.2012.731665

Weber, C. R., Lavine, H., Huddy, L., & Federico, C. M. (2014). Placing racial stereotypes in context: Social desirability and the politics of racial hostility. *American Journal of Political Science*, *58*, 63–78. https://doi.org/10.1111/ajps.12051

Westfall, P. H. (2014). Kurtosis as peakedness, 1905–2014. RIP. *The American Statistician*, *68*, 191–195. https://doi.org/10.1080/00031305.2014.917055

Williams, M. L., & Burnap, P. (2015). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, *56*, 211–238.

Xu, K., Nosek, B., & Greenwald, A. G. (2014). Data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*, e3.

Zeitzoff, T. (2017). How social media is changing conflict. *Journal of Conflict Resolution*, *61*, 1970–1991. https://doi.org/10.1177/0022002717721392